(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 936 565 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
18.08.1999 Bulletin 1999/33

(51) Int Cl.6: G06F 17/30

(21) Application number: 99103018.0

(22) Date of filing: 15.02.1999

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 16.02.1998 JP 3250398

(71) Applicant: Biomolecular Engineering Research Institute
Suita-shi, Osaka 565-0874 (JP)

(72) Inventor: Toh, Hiroyuki
Suita-shi, Osaka 565-0862 (JP)

(74) Representative: Howden, Christopher Andrew
FORRESTER & BOEHMERT
Franz-Joseph-Strasse 38
80801 München (DE)

(54) Method of searching database of three-dimensional protein structures

(57) A method of searching a database of three-dimensional protein structures. The method comprises the steps of setting a three-dimensional protein structure; forming a two-dimensional binary distance map based on the three-dimensional protein structure; forming a one-dimensional peripheral distribution based on the distance map; and comparing the one-dimensional peripheral distribution of a protein structure with that of another protein structure a dynamic programming algorithm. The method increases detection sensitivity and search speed.

Printed by Jouve, 75001 PARIS (FR)

D scription

BACKGROUND OF THE INVENTION

Field of the Invention:

[0001] The present invention relates to a method of searching a database of three-dimensional protein structures (hereinafter simply referred to as a "protein structure database"), and particularly to a method of searching a protein structure database through use of peripheral distributions of distance maps.

Description of the Related Art:

[0002] The three-dimensional structure of a protein provides various kinds of information in terms of pharmacology and physical chemistry, as well as important information in terms of biology. With recent progress in structure determination techniques, the number of entries in a protein structure database has increased drastically. One technique for analyzing proteins is comparison analysis in which similar-structures are compared to each other. Comparative analysis requires a technique for searching a structure database of huge size for structures resembling a three-dimensional structure obtained by a researcher.

SUMMARY OF THE INVENTION

[0003] In view of the foregoing, an object of the present invention is to provide a method of searching a protein structure database with peripheral distributions of distance maps, where a protein structure, which is three-dimensional information, is converted into one-dimensional information called peripheral distribution and then subjected to a dynamic programming algorithm (DP). The method can realize high speed search with high detection sensitivity.

[0004] In order to achieve the above object, the present invention provides a method of searching a database of three-dimensional protein structures, comprising the steps of setting a three-dimensional protein structure; forming a two-dimensional distance map based on the three-dimensional protein structure; forming a one-dimensional peripheral distribution based on the distance map; and comparing the one-dimensional peripheral distribution with that for another three-dimensional protein structure by use of a dynamic programming algorithm.

[0005] Preferably, the distance map is a two dimensional image and has a structure of a triangular matrix in which respective columns or respective rows correspond to respective residues of a protein;, the i-th row corresponds to the i-th amino acid residue counted from the N terminal end, and the j-th column corresponds to the j-th amino acid residue counted from the N terminal end; each element (i, j) of the matrix corresponds to the distance between the α carbon of the i-th residue and the α carbon of the j-th residue; and when the distance is smaller than or equal to a given threshold r_0, a dot is assigned to that portion, and when the distance is greater than the threshold r_0, a blank space is assigned to that portion, which operation is performed for each element in order to complete a binary distance map.

[0006] Preferably, the peripheral distribution is composed of a vertical peripheral distribution obtained in the form of a distribution of the frequency of dots at respective rows in a binary distance map and a horizontal peripheral distribution obtained in the form of a distribution of the frequency of dots at respective columns in the binary distance map.

[0007] Preferably, for comparison between peripheral distributions, an alignment score obtained by the dynamic programming algorithm is used as a similarity between corresponding protein structures.

[0008] A two-dimensional matrix, D, is required for the comparison of peripheral distributions. Each element of the matrix D is preferably obtained by solving the following recurrence equation:

$$D_{i,j} = \max \{D_{i-1,j-1} + s_{i,j}, D_{i-1,j} - g, D_{i,j-1} - g\}$$

where

$S_{i,j}$ indicates the similarity between the i-th element of the peripheral distribution of protein A and the j-th element of the peripheral distribution of protein B; and
g = 5 : gap penalty (however, g = 0 at the boundary)

[0009] Through the solution of the equation, the similarity is accumulated from the upper left corner toward the lower right corner of th matrix D, considering ins rtion and deletion. Then, the similarity between two peripheral distributions is obtained as a value for the element of the lower right corner of the matrix D.

[0010] $s_{i,j}$ is obtained by the following equation:

$$S_{i,j} = a/\{(N^A_i - N^B_j)^2+b\} + a/((C^A_i - C^B_j)^2+b)$$

where

N^A_j indicates the j-th frequency of the vertical peripheral distribution of protein A;
C^A_i indicates the i-th frequency of the horizontal distribution of protein A;
N^B_j indicates the j-th frequencies of the vertical peripheral distributions of protein B;
C^B_j indicates the j-th frequencies of the horizontal peripheral distribution of protein B; and where a = 50, and b = 2.

[0011] Preferably, a dot frequency R of a distance map is defined as follows:

R = number of dot elements in a distance map/

total number of elements in the distance map; and

the threshold is determined such that the dot frequency R falls within a predetermined range, and thus the detection sensitivity is increased.

[0012] More preferably, the threshold is determined such that the dot frequency R falls within the range of 0.12 to 0.16.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013]

FIG. 1 is a diagram showing the structure of a database search system according to an embodiment of the present invention;
FIG. 2 shows a flowchart illustrating a search performed through use of the protein structure database search system of FIG. 1;
FIG. 3 is an explanatory view showing a method of forming a binary distance map in accordance with the embodiment of the present invention;
FIGS. 4(a) and 4(b) are diagrams each showing a three-dimensional structure of thioredoxin;
FIG. 5 is a diagram showing distance maps of thioredoxin shown in FIGS. 4(a) and 4(b);
FIGS. 6(a) and 6(b) are explanatory views showing a method of forming a peripheral distribution;
FIG. 7 is an explanatory view showing a method of calculating the similarity between peripheral distributions according to the present invention;
FIG. 8 is a list of data used in measurement of the performance in a specific example of the present invention;
FIG. 9 is a diagram showing an effect of the dot frequency R on the search sensitivity;
FIG. 10 is an explanatory view showing a method of evaluating the detection sensitivity;
FIG. 11 is a table showing the effect of the dot frequency R on the detection sensitivity;
FIG. 12 is a table showing the result of comparison between the present search method and the search method utilizing the DDP;
FIGs. 13(a) and 13(b) show an example of a structure database search in which β-lactoglobulin is used as a query structure;
FIGS. 14(a) and 14(b) show the result of a search in which heat shock protein 70 (HSP 70) is used as a query structure; and
FIGS. 15(a) and 15(b) show the result of a search in which biotin carboxylase is used as a query structure.

DESCRIPTION OF THE PREFERRED EMBODIMENT

[0014] An embodiment of the present invention will next be described in detail with reference to the drawings.

[0015] FIG. 1 shows the structure of a protein structure database search system according to an embodiment of the present invention; and FIG. 2 shows a flowchart illustrating a search performed through use of the protein structure database search system of FIG. 1.

[0016] In FIG. 1, mineral 1 denotes an input section, which includes a parameter input section 2, a query structure-input section 3, and a list input section 4 for inputting a list of the file names of ntry files of a database. Numeral 10 denotes a processing section, which includes a memory (ROM) 11 and a data file section 12. The memory 11 stores therein a program for controlling the overall system. The data file section 12 stores therein the list of the file names of entry files of the database, data of the query structure, parameter values, and the protein structure database.

[0017] The processing section 10 further comprises a first distance map forming section 21, a first peripheral distribution forming section 22, an entry file read-in judgment section 23, a coordinate data input section 24, a second distance map forming section 25, a second peripheral distribution forming section 26, a similarity calculation section 27, a sort section 28, a search-result list output section 29, and a data output section 31.

[0018] The parameter value input section 2 is applied to input parameters a, b, and g for the DP, as well as a threshold r_0 shown in FIG. 3, which will be described later. The threshold r_0 is used in the distance map forming sections 21 and 25, while values of the parameters a, b, and g are used in the calculation section 27.

[0019] The query structure input section 3 reads in coordinates of a query structure to be searched. For example, when the structure of one of the proteins shown in FIG. 8 is selected as a query structure to be searched, the coordinates of the structure corresponding to the selected protein are input.

[0020] The protein structure database is not a single file but is composed of a plurality of independent files of data regarding individual protein structures. The list input section 4 reads in only a list of the file names.

[0021] The entry file read-in judgment section 23 judges whether the entire database has been read.

[0022] The coordinate data input section 24 successively reads in structure data from the files of the database in accordance with the list input by the input section 4; i.e., the coordinate data input section 24 reads in one file at a time from the database, each file including the structure(s) of a protein(s).

[0023] The second distance map forming section 25 forms a distance map in accordance with the read-in structure data. The second peripheral distribution forming section 26 forms a peripheral distribution in accordance with the thus-obtained distance map.

[0024] The similarity calculation section 27 calculates similarity through comparison in which a peripheral distribution of a query structure is compared with a peripheral distribution of an entry of the database by use of the DP (dynamic programming).

[0025] On the basis of the thus-calculated similarity, the sort section 28 determines the position of a presently-handled entry of the structure database within a search result list calculated up to the present. That is, the sort section 28 sorts searched entries in accordance with similarity to the query structure.

[0026] Upon completion of read-in of all data in the structure database and the above-described calculation within the loop, searched entries of the structure database have been sorted in accordance with similarity to the query structure. The search result list output section 29 outputs the thus-obtained search result list.

[0027] Although a detailed description will be given hereinafter, a search performed through use of the above-described protein structure database search system will be described with reference to FIG. 2.

(1) First, the protein structure database search system is initialized (step s1). In this step, parameters, such as the threshold r_0 and parameters a, b, and g of the DP, are input.
(2) Subsequently, a query structure is input (step S2).
(3) A distance map for the query structure is formed (step S3).
(4) A peripheral distribution for the query structure is formed (step S4).
(5) Subsequently, a list of file names of entry files of the database is input (step S5).
(6) Next, a check is made as to whether all of the entry files have been read in (step S6).
(7) When the result of the judgment in step S4 is NO, coordinate data are obtained from an entry file within the file name list (step S7).
(8) subsequently, a distance map is formed (step S8).
(9) Next, a peripheral distribution is formed (step S9).
(10) Subsequently, the similarity between a peripheral distribution of the query structure and the peripheral distribution of the database entry is calculated by means of the DP (step S10).
(11) Subsequently, sorting on the basis of the similarly is performed (step S11), and the processing proceeds back to step S4. The above-described procedure is repeated until all of the entry files are read in.
(12) When it is judged in step S4 that all of the entry files have been read in, a search result list is output (step S12).

[0028] Next, the method of searching the protein structure database will be described in detail.

[0029] In some techniques, the three—dimensional structure of a protein is converted into a distance map—which can be treated as a two-dimensional image—based on inter-residue distances and is displayed. As will be described later, when two proteins are similar in three-dimensional structure, their patterns on the respective distance maps are also similar to each other, even if their amino acid sequences differ. Accordingly, a protein having a similar structure can be found through comparison of distance maps.

[0030] Each distance map can be regarded or handled as a two-dimensional image. Pattern recognition of such a two-dimensional image is an important theme to be studied in the field of computer vision. In the present invention, among the methods used for the pattern recognition of two-dimensional images, a classical peripheral distribution scheme is used in order to covert a distanc map into one-dimensional information.

[0031] FIG. 3 shows a method of forming a binary distance map in accordance with the embodiment of the present invention.

[0032] The three-dimensional structure of a protein can be converted into a distance map, which is an two-dimensional image, through utilization of the distance between α carbons in residues thereof. In the present embodiment, a binary distance map is prepared in the following manner for conversion to a peripheral distribution.

[0033] The distance map has a structur of a triangular matrix, in which respective columns or respective rows correspond to respective residues of a protein. For example, the i-th row corresponds to the i-th amino acid residue counted from the N terminal end, and the j-th column corresponds to the j-th amino acid residue counted from the N terminal end. Each element (i, j) of the matrix corresponds to the distance between the α carbon of the i-th residue and the α carbon of the j-th residue. When the distance is smaller than or equal to a given threshold value (constant) r_0, a dot is assigned to that portion, and when the distance is greater than the threshold value r_0, a blank space is assigned to that portion. This operation is performed for each element in order to complete the distance map.

[0034] Next, there will be described comparison between distance maps.

[0035] FIG. 4(a) is a view showing a three-dimensional structure of thioredoxin derived from humans, whereas FIG. 4(b) is a view showing a three-dimensional structure of thioredoxin derived from bacteria.

[0036] As shown in FIGS. 4(a) and 4(b), the human thioredoxin and the bacteria thioredoxin have similar three-dimensional structures, although their amino acid sequence identity is only 23.3%.

[0037] FIG. 5 shows distance maps which correspond to the two structures shown in FIGS. 4(a) and 4(b) and which are formed through the steps of FIG. 3. As shown in FIG. 4, the two structures are similar to each other, although the sequence identity is only 23.3%. Reflecting the structural similarity, the distance maps are similar to each other. Accordingly, similarity between the two structures is expected to be evaluated not through comparison of their three-dimensional structures but through comparison of patterns on their distance maps.

[0038] Next, formation of peripheral distribution will be described.

[0039] FIG. 6 is an explanatory view showing a method of forming a peripheral distribution.

[0040] First, a method of forming a peripheral distribution used in the field of character recognition is described.

[0041] Consider that a letter "A" is drawn on a plane. The plane is divided into small squares by a mesh, and each square is colored black or white (coded in binary form) in order to represent the letter "A." For each row, the black lements are counted so as to obtain a frequency of black elements for the row. This procedure is repeated for all the rows in order to obtain a vertical peripheral distribution V.

[0042] A similar procedure is performed in order to obtain a horizontal peripheral distribution H. For example, the frequency of the third row in the vertical distribution is 3, since three black elements are present in the third row in the matrix of FIG. 6(a). The vertical and horizontal peripheral distributions are considered to represent the feature of the character "A." In the field of character recognition for printed Chinese characters, characters are recognized on the basis of such peripheral distributions.

[0043] Since a distance map can be regarded as a two-dimensional image, vertical and horizontal peripheral distributions can be formed for the distance map according to a method similar to the method described above. FIG. 6(b) shows vertical and horizontal peripheral distributions formed by such a method. Via the distance map, which is two-dimensional information, the three-dimensional structure of a protein, which is three-dimensional information, can be converted into peripheral distributions, which are one-dimensional information.

[0044] As described above, the peripheral distributions of-a distance map are considered to represent the characters of the distance map. Therefore, a similar structure can be recognized through comparison of peripheral distributions.

[0045] Next, there will be described methods of calculating the similarity between peripheral distributions.

[0046] First, there will be described a first method of calculating the similarity between peripheral distributions.

[0047] Character recognition is performed on the basis of similarity between peripheral distributions. The similarity is calculated through simple superimposition of the distributions or correlation of the Fourier spectrums of the distributions. However, neither method can deal with insertion or deletion, which occurs in proteins, but is not considered in ordinary character recognition.

[0048] In the technique for comparison of sequence data, an alignment score obtained as a result of a DP matching has been used as a similarity in which insertion and deletion are taken into consideration. Since peripheral distributions, like sequence data, are one-dimensional information, the present inventor tried to apply DP matching to peripheral distributions in a manner shown in FIG. 7.

[0049] FIG. 7 is an explanatory view showing a method of calculating the similarity between peripheral distributions according to the present invention.

[0050] In the DP for comparison of one-dimensional data, as shown in FIG. 7, a two-dimensional matrix, D, is required for the comparison of peripheral distributions. Each element of the matrix D is obtained by solving the following recurrence equation:

$$D_{i,j} = \max\{\text{①}D_{i-1,j-1} + S_{i,j}, \text{②}D_{i-1,j} - g, \text{③}D_{i,j-1} - g\}$$

where

$S_{i,j}$ indicates the similarity between the i-th element of the peripheral distribution of protein A and the j-th element of the peripheral distribution of protein B; and
g = 5 : gap penalty (however, g = 0 at the boundary)

[0051] Through the solution of the equation, the similarity is accumulated from the upper left corner toward the lower right corner of the matrix D, considering insertion and deletion. Then, the similarity between two peripheral distributions is obtained as a value for the element of the lower right corner of the matrix D.

[0052] $s_{i,j}$ is obtained by the following equation:

$$S_{i,j} = a/\{(N^A_i - N^B_j)^2 + b\} + a/\{(C^A_i - C^B_j)^2 + b\}$$

where

N^A_i indicates the j-th frequency of the vertical peripheral distribution of protein A;
C^A_i indicates the i-th frequency of the horizontal distribution of protein A;
N^B_j indicates the j-th frequencies of the vertical peripheral distributions of protein B;
C^B_j indicates the j-th frequencies of the horizontal peripheral distribution of protein B; and where a = 50, and b = 2.
$S_{i,j}$ indicates the sum of the similarity in frequency of vertical distribution and the similarity in frequency of horizontal distributions between the i-th residue of protein A and the j-th residue of protein B. The similarity between two peripheral distributions is proportional to the size of the proteins corresponding to the distributions, even when the structures under comparison are not similar to each other. To eliminate the size dependency, the similarity is divided by the length of aligned peripheral distributions by DP matching, which is used as the similarity between two structures.

[0053] Next, a specific example will be described.

[0054] The program is made by a program language, ANSI C.

[0055] The performance of the system was evaluated on a computer, DEC Alpha server $2100^{5/250}$. Protein Data Bank release #81 was used as the database for performance check, which is hereafter referred to as PDB.

[0056] Next, there is described data used in measurement of performance.

[0057] FIG. 8 shows data used in the measurement of performance in the specific example.

[0058] In order to investigate the sensitivity of the database search according to the present invention, the database search was performed with nine proteins having different structures in accordance with the method of the present invention, and the calculation time and detection sensitivity were measured.

[0059] The nine proteins having different structures were selected as follows.

[0060] First, in order to prevent the performance measurement for the method of the present invention from depending on the kinds of structures, three-dimensional structures were selected from each of three representative classes; i.e., mainly α, mainly β and α/β.

[0061] Three kinds of proteins having different structures (categorized in different super families in accordance with the SCOP classification) were selected from each structural class, and search was performed. Then, nine proteins were used as query structures. FIG. 8 shows a list of the thus-selected nine proteins.

[0062] For comparison with the method of the present invention, database search with a double dynamic programming (DDP) method, Which is a more precise structural comparison, as well as database search at the level of amino acid sequence, were performed with the nine proteins.

[0063] However, the search with the DDP was not performed for proteins having a size greater than 200 residues, because they required an excessively long time. The structural comparison with the DDP is disclosed in detail in Japanese Patent Application No. 8-340727, which was filed by the present inventor.

[0064] The database search at the level of amino acid sequence was performed with a program FASTA available at the internet site GenomeNet. Since this search was performed on a different computer, the calculation time is not shown.

[0065] Next, there will be described a dot frequency R, which is a factor for determining the detection sensitivity.

[0066] FIG. 9 shows an effect of the dot frequency R on the detection sensitivity.

[0067] An attempt was made to set the threshold r_0 (see FIG. 3) for obtaining a binary distance map to an optimal value for database search.' However, since the threshold r_0 for optimizing the detection sensitivity varied from protein to protein, the threshold r_0 could not be fixed to one value. The present inventor considered that, sinc the threshold r_0 is a factor that determines the frequency of dots in a distance map, th detection s nsitivity is affected not by the threshold r_0 itself but by the dot frequency that is considered to relate to the characterization of the pattern of the distance map. Thus, the present inventor investigated the relationship between the dot frequency R and the detection sensitivity, while defining the dot frequency R as follows:

$$R = \text{number of black elements of a distance map/ total number of elements of the distance map.}$$

[0068] As shown in FIG. 9, peripheral distribution characterizes the distance map or the tertiary structure when the dot frequency R is excessively small or excessively large. Therefore, a proper value for the dot frequency R must be found.

[0069] The analysis described above reveals that a high detection sensitivity is obtained when the threshold r_0 is determined such that the dot frequency R falls within the range of 0.12 to 0.16.

[0070] The threshold r_0 that causes the dot frequency R to fall within the range described above varies from protein to protein.

[0071] FIG. 10 is an explanatory view to show a method of evaluating the detection accuracy.

[0072] In FIG. 10, symbol A denotes proteins that are categorized in a family in the SCOP (structure classification database) to which a query structure belongs; symbol B denotes proteins that are not categorized in the same family in the SCOP but are categorized in a class (superfamily) for proteins having structures that share the same topology in the structure with the query, but have weak similarity in amino acid sequence to the query; and proteins C (no symbol) are proteins classified into different superfamilies, to which a query structure does not belong.

[0073] As shown in FIG. 10, the names of entries in the database are output as a result of the search, in the form of a list where the entries are sorted in descending order of similarity to a query structure. In the test, it is needless to say that the entry at the top of the output list corresponds to the query structure itself, since each query structure is obtained from the entries of the structure database.

[0074] In the list, the members of class A or class B are regarded as "success." This process is repeated from the top of the list until a protein categorized in the class C is first found. The number (L) of proteins in the class A and the number (M) of proteins in the class B contained in the run of success are counted in order to calculate the ratio of the number L to the entire number of the class-A proteins of the structure.database and the ratio of the number M to the entire maker of the class-B proteins of the structure database. These ratios were used as indictors for detection sensitivity. Note that a class A or class B protein in the list is not counted for L or M if they do not belong to the run of success.

[0075] FIG. 11 is a table showing the effect of the dot frequency R on the detection sensitivity.

[0076] The first column shows the names of query structures. The fifth column shows the number of entries in the structure database classified into the classes A and B. The second to fourth columns respectively show the values of L and M for respective ranges for R. In each of the second to fifth columns, the number L of A-class proteins is shown on the left side of the "+" symbol, and the number M of B-class proteins is shown on the right side of the "+" symbol. When the value of R is less than 0.12 (second column), the sensitivity in detecting class-B proteins is extremely low, although most of the class-A proteins are detected for each query protein.

[0077] When the value of R is greater than 0.16 (fourth column), the sensitivity in detecting class-B proteins drops for some proteins, although class-A proteins are detected with high sensitivity. In contrast, when the value of R is greater than or equal to 0.12 and less than or equal to 0.16 (third column), the sensitivity in detecting class-B proteins is high.

[0078] The effect of the dot frequency R is shown for each of three structural classes; i.e., mainly α, mainly β, and α/β.

[0079] In FIG. 12, in order to demonstrate the performance of the method of the present invention, the result of database search according to the method of the present invention is compared with the result of database search according to the DDP previously proposed by the present inventor.

[0080] Although structure comparison performed by the DDP is more precise than the method of the present invention, it takes a huge amount of time for calculation. As shown in FIG. 12, for class-A proteins, the search method of the present invention provides a detection sensitivity substantially equal to that obtained in the case of the search method utilizing the DDP. However, for class-B proteins, the search method of the present invention provides a detection sensitivity higher than that obtained in the case of the search method utilizing the DDP. Despite the higher sensitivity of current invention, the calculation time is greatly shortened compared to the case of the search method With the DDP. This demonstrates the superiority of the method of the present invention over the search method with the DDP, although the number of compared samples is small.

[0081] FIGS. 13(a) and 13(b) show an example of a structure database search in which β-lactoglobulin is used as a query structure. In FIGS. 13(a) and 13(b), a bar chart of frequency distribution of similarity is shown on the left side, an output list is shown on the right side.

[0082] β-lactoglobulin has a β-barrel structure with eight β strands and belongs to the lipocalin family. In the SOOP, the lipocalin and a family of proteins having a β-barrel structure with ten β strands form a superfamily in terms of structure.

[0083] In the present invention, class A is defined as the lipocalin family, and class B is defined as proteins with a β -barrel structure composed of ten β strands.

[0084] FIG. 13(a) shows the result of a search with the DDP. As shown in FIG. 12, the search with the DDP could not detect class-B proteins at all, although it could detect all of class-A proteins.

[0085] In contrast, as shown in FIG. 13(b), the search method of the present invention d tected may class-B proteins after detection of all the class-A proteins. Although only the top fifty proteins are output, class-B proteins were detected after the run of success.

[0086] FIGS. 14(a) and 14(b) show the result of a search in which heat shock protein 70 (HSP 70) is used as a query structure.

[0087] In this study, HSP 70 forms a family, which is used as class A. In the SCOP, actin and hexokinase are included in its classification for superfamily level. These were defined to form class B. since HSP 70 is a very large protein of about 400 residues in length, search with the DDP was difficult from the viewpoint of computation time.

[0088] Therefore, instead of the DDP, the FASTA of the GenomeNet was used for database search of HSP 70 at sequence level. FIG. 14(a) shows the result of search with the FASTA.

[0089] The FASTA could not detect any class-B proteins at all, although it could detect all of class-A proteins. In contrast, FIG. 14(b) shows the result of the search with the method of the present invention. As shown in FIG. 14(b), actin belonging to the class B was detected after detection of all the class-A proteins. However, hexokinase was not detected. The result of the FASTA demonstrates that, no significant similarity is observed at the sequence level, although HSP 70 and actin resemble each other in structure.

[0090] FIGS. 15(a) and 15(b) show the result of a search in which biotin carboxylase is used as a query structure.

[0091] Biotin carboxylases form one family by themselves, and therefore it was used as class A.

[0092] Although biotin carboxylase exhibits structural and functional similarity with D-Ala-D-Ala ligase and glutathione synthetase, no significant similarity is observed at the sequence level. Therefore, these were used as class B. Instead of the DDP, the FASTA of the GenomeNet was used for search and comparison, since biotin carboxylase is also a very large protein of about 400 residues in length. FIG. 15(a) shows the result of search with the FASTA. In this case, the FASTA could detect D-Ala-D-Ala ligases, as well as all of the class A proteins. However, no glutathione synthetase was detected. In contrast, as shown in PIG. 15(b), the method of the present invention could detect glutathione synthetases after detection of class A proteins and D-Ala-D-Ala ligase. Further, many glutathione synthetases are found after the run of success in the output list. However, they are not taken into consideration in the evaluation method described above.

[0093] As is apparent from the specific example, the detection method of the present invention has a higher detection sensitivity than the DDP and FASTA.

[0094] The present invention is not limited to the embodiments described above. Numerous modifications and variations of the present invention are possible in light of the spirit of the present invention, and they are not excluded from the scope of the present invention.

[0095] As described above, according to the present invention, the three-dimensional structure of a protein, which is three-dimensional information, is converted into a peripheral distribution, which is one-dimensional information, and is then subjected to comparison with a dynamic programming algorithm. Therefore, the detection sensitivity can be increased, and high speed search can be realized.

[0096] Thus, a database search with high speed and high sensitivity was realized, which would cope with rapid increase of the entry of protein structure database to make enormous contribution to biology, pharmacology and physical chemistry.

Claims

1. A method of searching a database of three-dimensional protein structures, comprising the steps of:

   (a) setting a three-dimensional protein structure;
   (b) forming a two-dimensional binary distance map based on the three-dimensional protein structure;
   (c) forming a one-dimensional peripheral distribution based on the binary distance map; and
   (d) comparing the one-dimensional peripheral distribution with that for another three-dimensional protein structure by a dynamic programming algorithm.

2. A method of searching a database of three-dimensional protein structures according to Claim 1, wherein said distance map is a two dimensional image and has a structure of a triangular matrix in which r spective columns or respective rows correspond to respectiv residues of a protein; the i-th row corresponds to the i-th amino acid residue counted from th N terminal end, and the j-th column corresponds to the j-th amino acid residue counted

from the N terminal end; each element (i, j) of the matrix corresponds to the distance between the α carbon of the i-th residue and the α carbon of the j-th residue; and when the distance is smaller than or equal to a given threshold r_0, a dot is assigned to that portion, and when the distance is greater than the threshold r_0, a blank space is assigned to that portion, which operation is performed for each element in order to complete the binary distance map.

3. A method of searching a database of three-dimensional protein structures according to Claim 2, wherein said peripheral distribution is composed of a vertical peripheral distribution obtained as a distribution of frequencies of dots at respective rows in a binary distance map and a horizontal peripheral distribution obtained as a distribution of frequencies of dots at respective columns in the binary distance map.

4. A method of searching a database of three-dimensional protein structures according to claim 3, wherein for comparison between peripheral distributions, an alignment score obtained by the dymamic programming algorithm divided by the alignment length is used as a similarity between two structures.

5. A method of searching a database of three-dimensional protein structures according to Claim 3, wherein a two dimensional matrix, D, is used for the comparison of peripheral distributions; each element of the matrix D is obtained by solving the following recurrence equation; through the solution of the equation, the similarity is accumulated from the upper left corner toward the lower right corner of the matrix D, considering insertion and deletion; and then, the similarity between two peripheral distributions is obtained as a value for the element of the lower right of the matrix D:

$$D_{i,j} = \max \{D_{i-1,j-1} + s_{i,j}, D_{i-1,j} - g, D_{i,j-1} - g\}$$

where

$g = 5$: gap penalty (however, $g = 0$ at the boundary),

and

$S_{I,j}$ is represented by the following equation and indicates the similarity between the i-th element of the peripheral distribution of protein A and the j-th element of the peripheral distribution of protein B:

$$S_{I,j} = a/ \{(N_I^A - N_j^B)^2 + b\} + a/ \{(C_i^A - C_j^B)^2 + b\}$$

where

N^A_j indicates the j-th frequency of the vertical peripheral distribution of protein A;
C^A_i indicates the i-th frequency of the horizontal distribution of protein A;
N^B_j indicates the j-th frequencies of the vertical peripheral distributions of protein B;
C^B_j indicates the j-th frequencies of the horizontal peripheral distribution of protein B; and
a and b are constants.

6. A method of searching a database of three-dimensional protein structures according to Claim 3, wherein a dot frequency R in the distance map is defined as follows:

R = number of black elements of a distance map/ total number of elements of the distance map;

and the threshold is determined such that the dot frequency R falls within a predetermined range, and the detection sensitivity is increased.

7. A method of searching a database of three-dimensional protein structures according to Claim 3, wherein the threshold is determined such that the dot frequency R falls within th range of 0.12 to 0.16.

F I G. 1

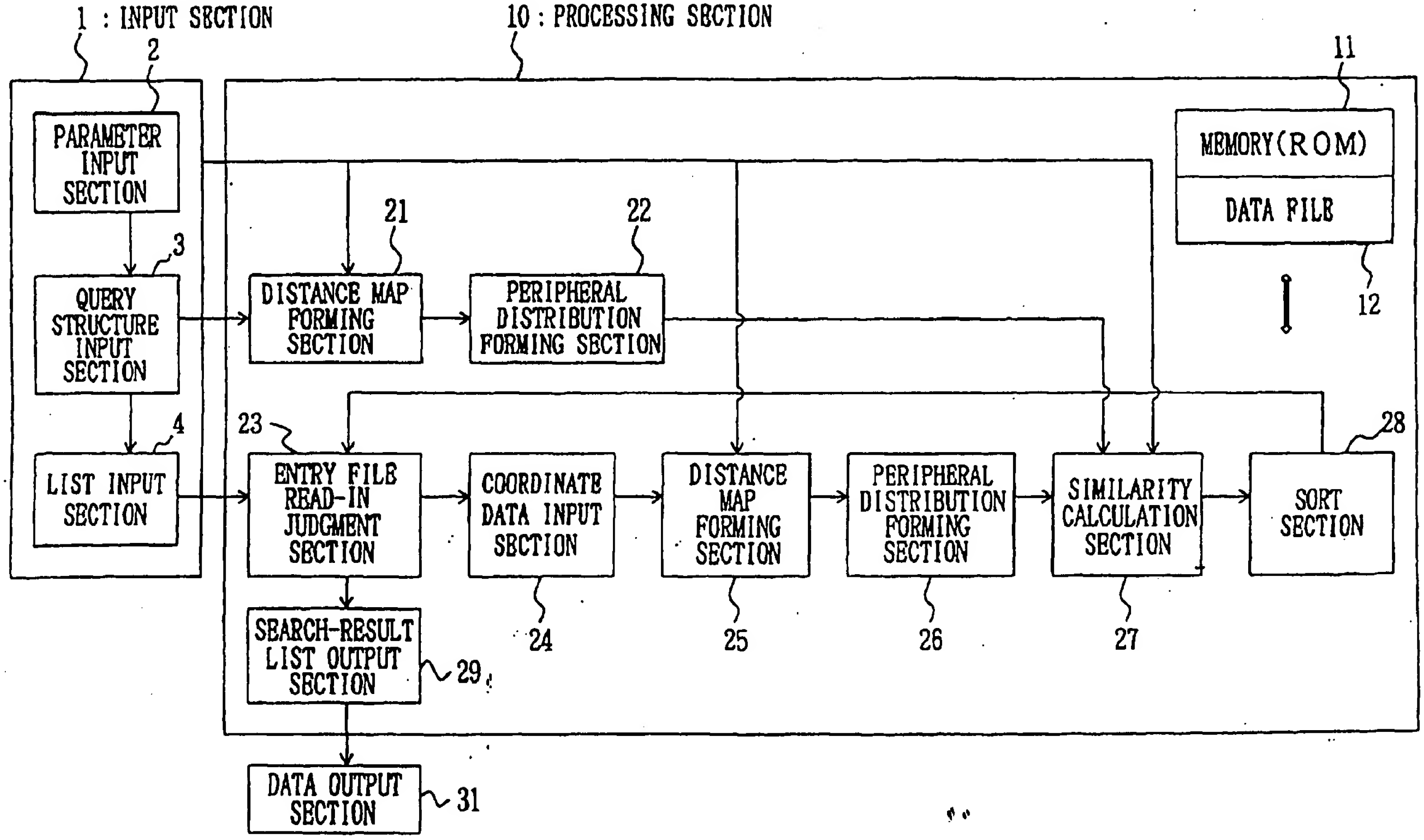

1 : INPUT SECTION
10 : PROCESSING SECTION
2
PARAMETER INPUT SECTION
3
QUERY STRUCTURE INPUT SECTION
4
LIST INPUT SECTION
21
DISTANCE MAP FORMING SECTION
22
PERIPHERAL DISTRIBUTION FORMING SECTION
11
MEMORY(ROM)
DATA FILE
12
23
ENTRY FILE READ-IN JUDGMENT SECTION
COORDINATE DATA INPUT SECTION
24
DISTANCE MAP FORMING SECTION
25
PERIPHERAL DISTRIBUTION FORMING SECTION
26
SIMILARITY CALCULATION SECTION
27
SORT SECTION
28
SEARCH-RESULT LIST OUTPUT SECTION
29
DATA OUTPUT SECTION
31

FIG. 2

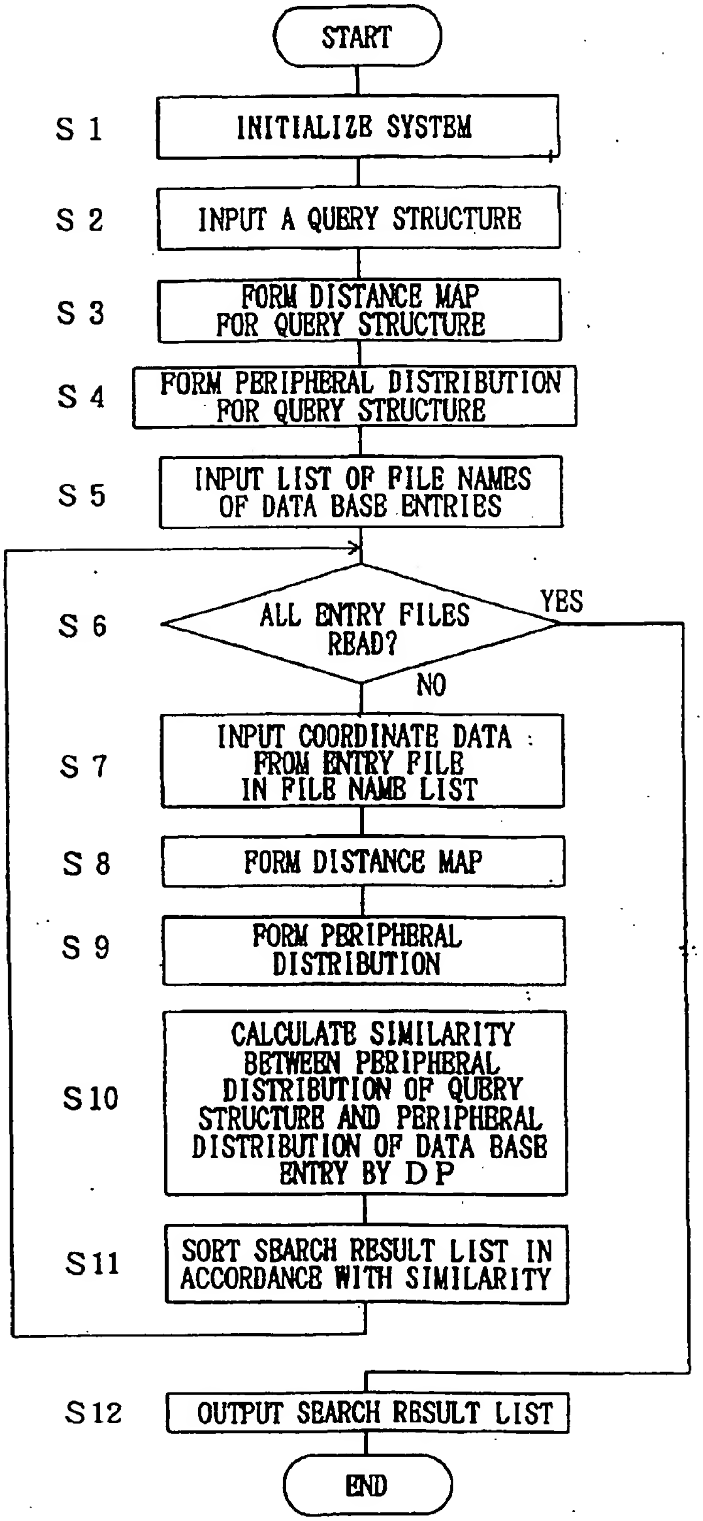

START
S 1
INITIALIZE SYSTEM
S 2
INPUT A QUERY STRUCTURE
S 3
FORM DISTANCE MAP FOR QUERY STRUCTURE
S 4
FORM PERIPHERAL DISTRIBUTION FOR QUERY STRUCTURE
S 5
INPUT LIST OF FILE NAMES OF DATA BASE ENTRIES
S 6
ALL ENTRY FILES READ?
YES
NO
S 7
INPUT COORDINATE DATA FROM ENTRY FILE IN FILE NAME LIST
S 8
FORM DISTANCE MAP
S 9
FORM PERIPHERAL DISTRIBUTION
S10
CALCULATE SIMILARITY BETWEEN PERIPHERAL DISTRIBUTION OF QUERY STRUCTURE AND PERIPHERAL DISTRIBUTION OF DATA BASE ENTRY BY DP
S11
SORT SEARCH RESULT LIST IN ACCORDANCE WITH SIMILARITY
S12
OUTPUT SEARCH RESULT LIST
END

FIG. 3

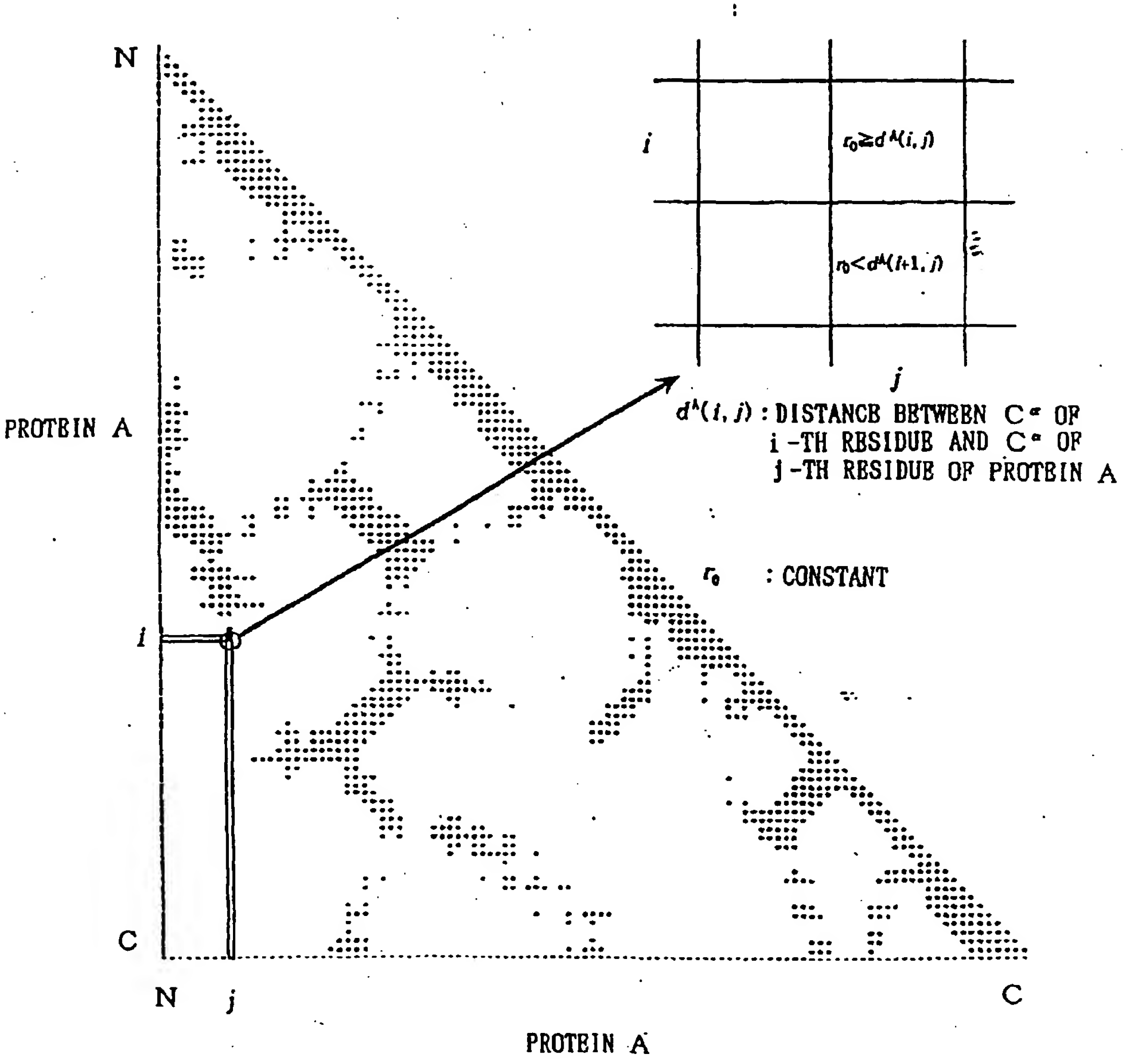
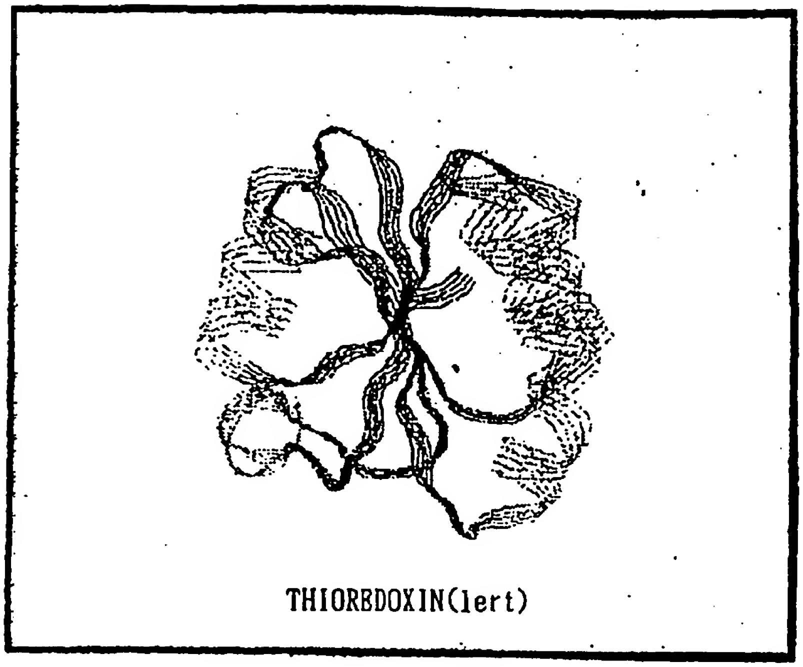

N
PROTEIN A
i
C
N
j
C
PROTEIN A
i
$r_0 \geq d^A(i, j)$
$r_0 < d^A(i+1, j)$
j
$d^A(i, j)$: DISTANCE BETWEEN C^α OF i-TH RESIDUE AND C^α OF j-TH RESIDUE OF PROTEIN A
r_0 : CONSTANT

FIG. 4 (a)



SBQUENCB IDBNTITY : 23.3%

FIG. 4 (b)

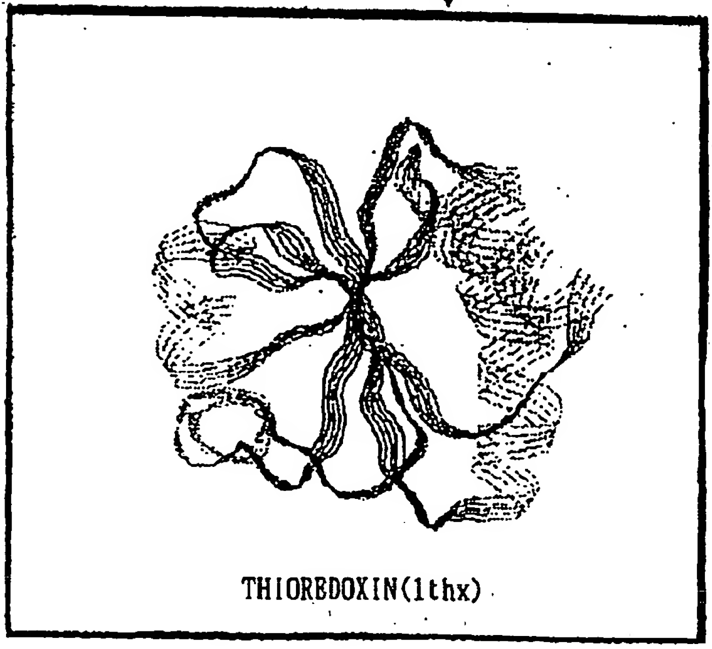


FIG. 5

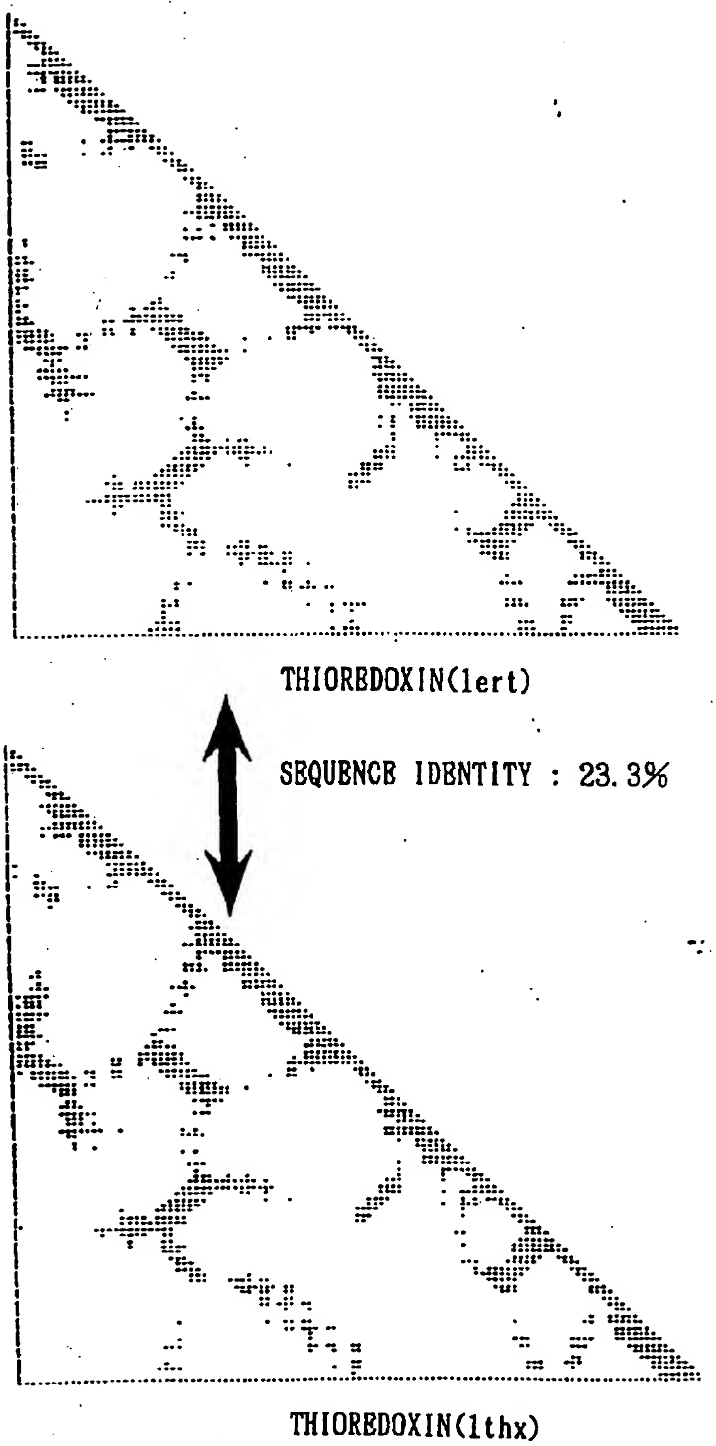


FIG. 6 (a)

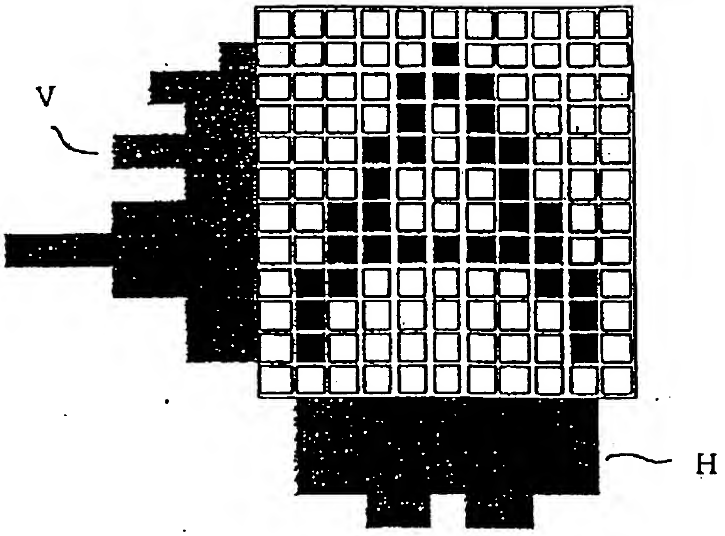
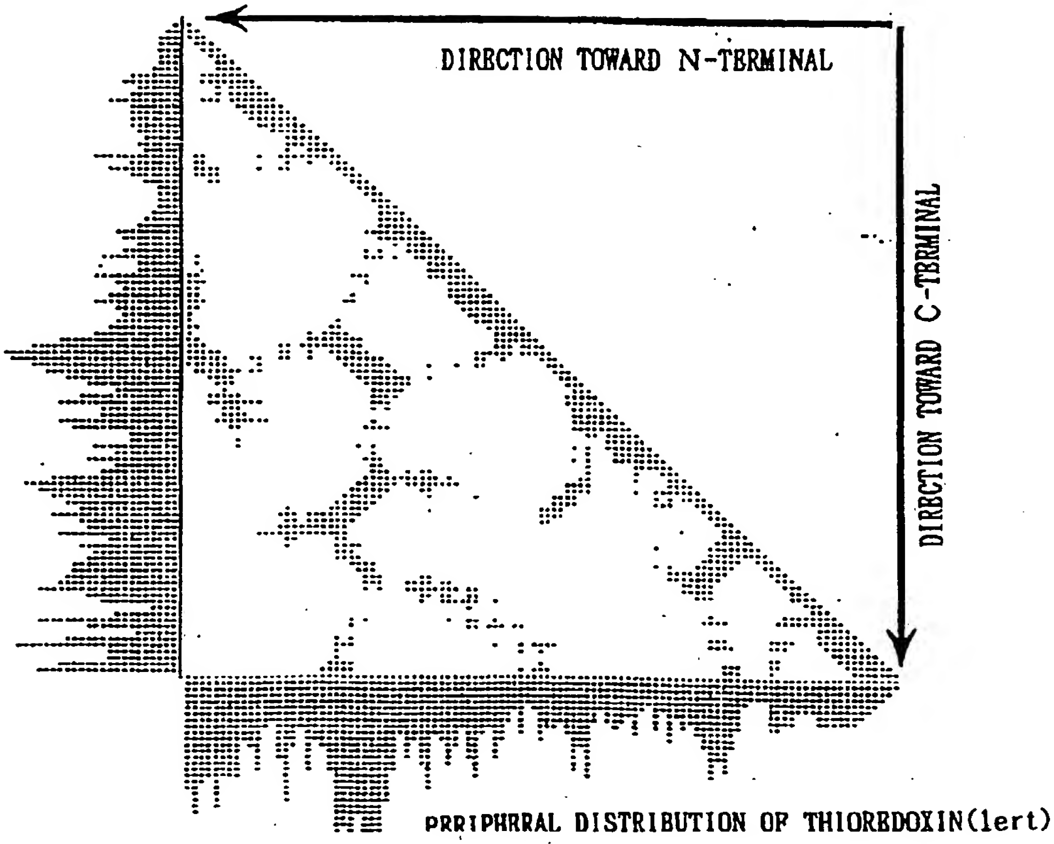
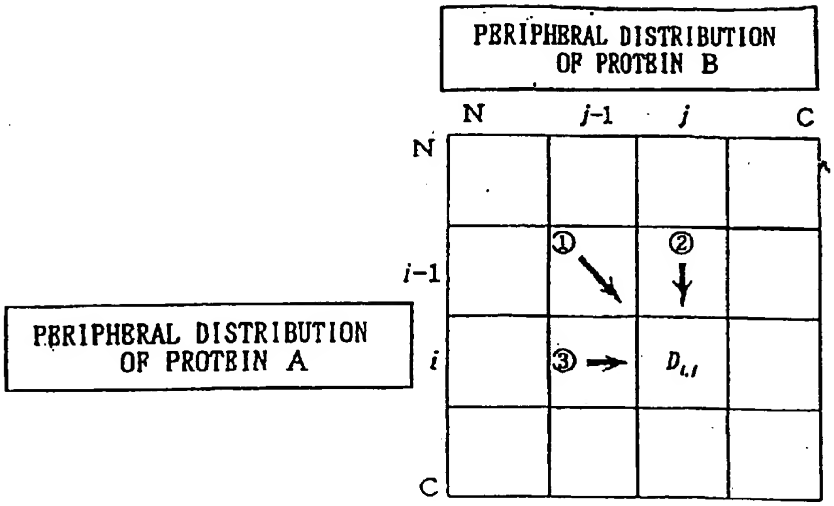

V
H

FIG. 6 (b)


DIRECTION TOWARD N-TERMINAL
DIRECTION TOWARD C-TERMINAL
PERIPHERAL DISTRIBUTION OF THIOREDOXIN(1ert)

F I G. 7


PERIPHERAL DISTRIBUTION OF PROTEIN B
N
j-1
j
C
N
i-1
i
C
PERIPHERAL DISTRIBUTION OF PROTEIN A
①
②
③
$D_{i,j}$

comparison matrix D

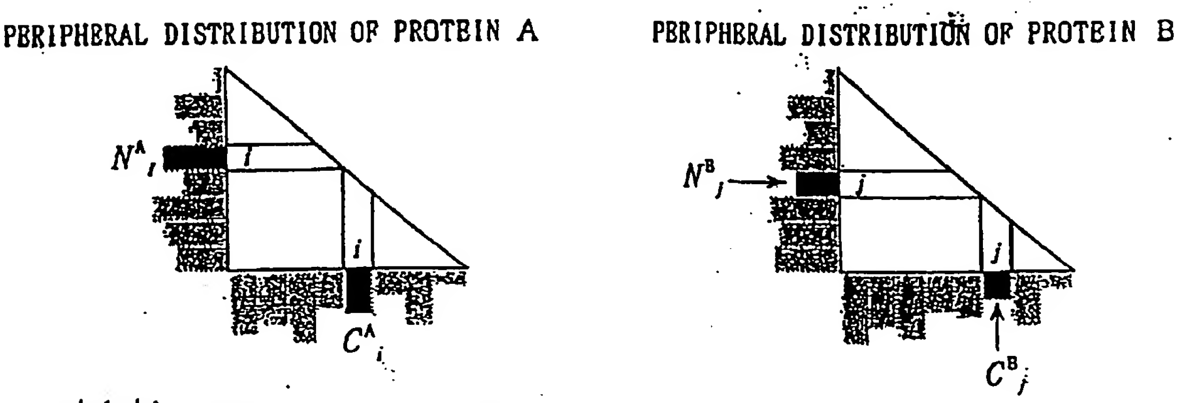
$$D_{i,j} = \max\{\textcircled{1} D_{i-1,j-1} + s_{i,j},\ \textcircled{2} D_{i-1,j} - g,\ \textcircled{3}\ D_{i,j-1} - g\}$$

$g=5$: GAP PENALTY (HOWEVER, g = 0 AT BOUNDARY)

$s_{i,j}$: SIMILARITY BETWEEN i-TH FREQUENCY OF PERIPHERAL DISTRIBUTION OF PROTEIN A AND j-TH FREQUENCY OF PERIPHERAL DISTRIBUTION OF PROTEIN B

$$s_{i,j} = a / \{(N^A_i - N^B_j)^2 + b\} + a / \{(C^A_i - C^B_j)^2 + b\}$$

$a = 50,\ b = 2$

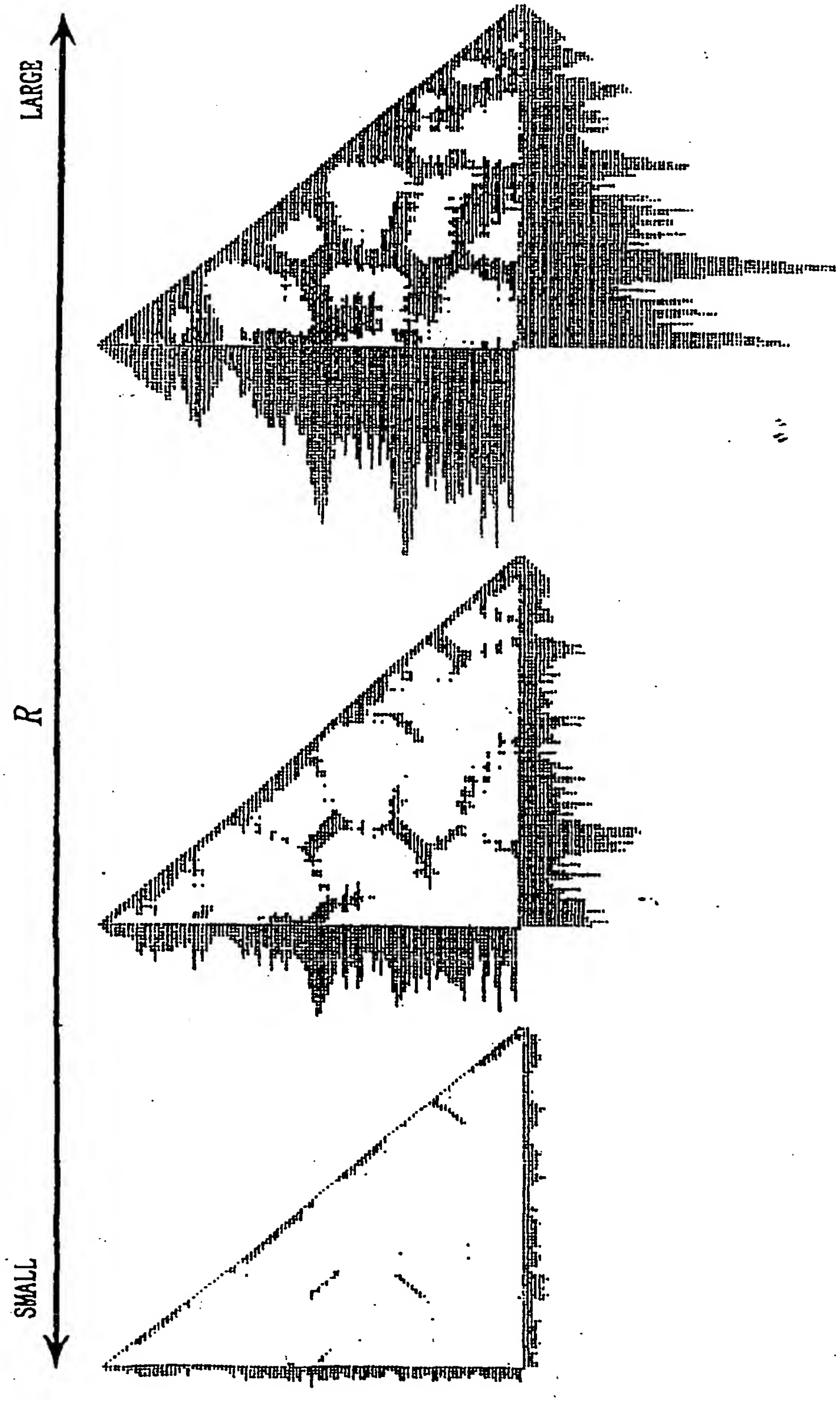

PERIPHERAL DISTRIBUTION OF PROTEIN A
N^A_i
i
i
C^A_i
PERIPHERAL DISTRIBUTION OF PROTEIN B
N^B_j
j
j
C^B_j

$N^A_i(C^A_i)$: FREQUENCY OF PERIPHERAL DISTRIBUTION IN N(C) TERMINAL DIRECTION CORRESPONDING TO THE i-TH RESIDUE OF PROTEIN A

FIG. 8

PROTEIN	CODE	NUMBER OF RESIDUES
mainly α		
1. myoglobin	1mbc	153
2. cytochrome c	1ccr	112
3. cytochrome P450	1oxa	403
mainly β		
4. β-lactoglobulin	1beb(A CHAIN)	162
5. telokin-like protein	1tul	108
6. interleukin-1 β	1ilb	153
α/β		
7. biotin carboxylase	1bnc	449
8. heat-shock protein 70kDa (HSP70)	1atr	386
9. thioredoxin	1ert	105

FIG. 9



F I G. 10

LIST OF DETECTED PROTEINS

```
     1th pdb1atr.ent : HEAT-SHOCK COGNATE 70 KD PROTEIN (44 KD ATPASE N-TERMINAL  (1)
     2th pdb1ngf.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
     3th pdb1ngj.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
     4th pdb3hsc.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
     5th pdb1ate.ent : HEAT-SHOCK COGNATE 70 KD PROTEIN (44 KD ATPASE N-TERMINAL  (1)
     6th pdb1nga.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
A    7th pdb1ngh.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)      L
     8th pdb1ngg.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
     9th pdb1nge.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
    10th pdb1ngb.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
    11th pdb1ngi.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
    12th pdb1ngc.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
    13th pdb1ngd.ent : HEAT-SHOCK COGNATE 70KD PROTEIN (44KD ATPASE N-TERMINAL  (1)
    14th pdb1kax.ent : 70KD HEAT SHOCK COGNATE PROTEIN ATPASE DOMAIN, K71M MUTANT  (1)
B   15th pdb1hpm.ent : 44K ATPASE FRAGMENT (N-TERMINAL) OF 70KDA HEAT-SHOCK COGNATE  (1)   M
    16th pdb1kay.ent : 70KD HEAT SHOCK COGNATE PROTEIN ATPASE DOMAIN, K71A MUTANT  (1)
    17th pdb1kaz.ent : 70KD HEAT SHOCK COGNATE PROTEIN ATPASE DOMAIN, K71E MUTANT  (1)
C   18th pdb1atn.ent : DEOXYRIBONUCLEASE I COMPLEX WITH ACTIN  (A)
    19th pdb2btf.ent : BETA-ACTIN-PROFILIN COMPLEX  (A)
    20th pdb1glk.ent : GLUCOKINASE (ATP:D-HEXOSE 6-PHOSPHOTRANSFERASE)  (1)
A   21th pdb4gpd.ent : APO-D-GYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (3)
    22th pdb1tad.ent : TRANSDUCIN-ALPHA (GT-ALPHA-GDP-ALF, T-ALPHA-GDP-ALF)  (B)
    23th pdb1nlg.ent : OXIDIZED NADP-LINKED GLYCERALDEHYDE-3-PHOSPHATE  (1)
    24th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (O)
    25th pdb4gpd.ent : APO-D-GYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (1)
B   26th pdb4gpd.ent : APO-D-GYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (4)
    27th pdb4gpd.ent : APO-D-GYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (2)
    28th pdb1pfk.ent : PHOSPHOFRUCTOKINASE (E.C.2.7.1.11) (R-STATE) COMPLEX WITH  (B)
    29th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (P)
    30th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (R)
    31th pdb1hdg.ent : HOLO-D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (Q)
    32th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (A)
    33th pdb1gyp.ent :  MOLECULE: GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE;  (B)
    34th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (O)
    35th pdb1gyp.ent :  MOLECULE: GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE;  (A)
    36th pdb1gyp.ent :  MOLECULE: GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE;  (C)
    37th pdb1gga.ent : D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (HOLO FORM)  (R)
    38th pdb1tad.ent : TRANSDUCIN-ALPHA (GT-ALPHA-GDP-ALF, T-ALPHA-GDP-ALF)  (A)
    39th pdb1tag.ent : TRANSDUCIN-ALPHA COMPLEXED WITH GDP AND MAGNESIUM  (1)
    40th pdb1gyp.ent :  MOLECULE: GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE;  (D)
    41th pdb1nlh.ent : REDUCED NADP-LINKED GLYCERALDEHYDE-3-PHOSPHATE  (1)
    42th pdb1pfk.ent : PHOSPHOFRUCTOKINASE (E.C.2.7.1.11) (R-STATE) COMPLEX WITH  (A)
    43th pdb1tad.ent : TRANSDUCIN-ALPHA (GT-ALPHA-GDP-ALF, T-ALPHA-GDP-ALF)  (C)
    44th pdb1hdg.ent : HOLO-D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE  (O)
    45th pdb1tnd.ent : TRANSDUCIN (ALPHA SUBUNIT) COMPLEXED WITH THE  (B)
    46th pdb6pfk.ent : PHOSPHOFRUCTOKINASE, INHIBITED T-STATE  (C)
    47th pdb1tnd.ent : TRANSDUCIN (ALPHA SUBUNIT) COMPLEXED WITH THE  (A)
    48th pdb1cer.ent :  MOLECULE: HOLO-D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE;  (P)
    49th pdb1tdf.ent : THIOREDOXIN REDUCTASE (E.C.1.6.4.5) MUTANT WITH CYS 138  (1)
    50th pdb4dbv.ent : GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE MUTANT WITH LEU 33  (O)
```

A PROTEINS CLASSIFIED IN THE SAME FAMILY IN P I R

B PROTEINS THAT ARE NOT CLASSIFIED IN THE SAME FAMILY IN P I R BUT CLASSIFIED IN THE SAME FAMILY IN S C O P

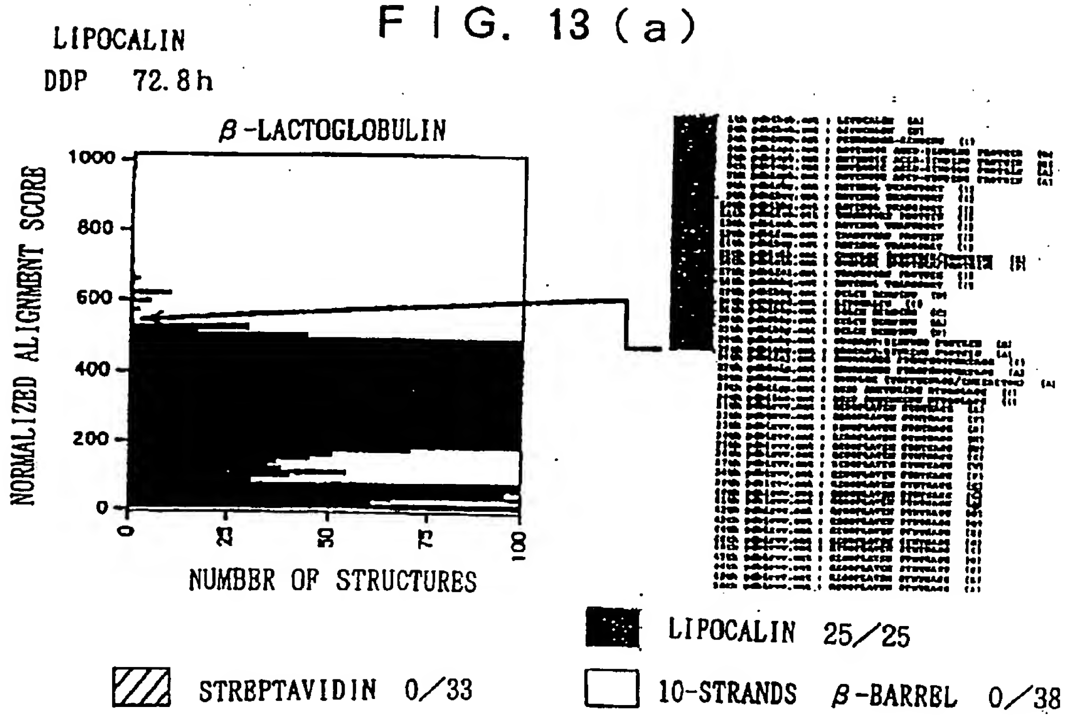
F I G. 11

PROTEIN	R<0.12	0.12≤R≤0.16	0.16<R	FAMILY+SIMILAR STRUCTURE
mainly α				
myoglobin (1mbc)	324 + 0	329 + 0	329 + 0	329 + 6
cytochrome c (1ccr)	44 + 0	53 + 0	52 + 0	72 + 0
cytochrome P450 (1oxa)	31 + 0	31 + 0	31 + 0	31 + 0
mainly β				
β-lactoglobulin (1beb)	25 + 0	25 + 25	25 + 1	25 + 71
telokin-like protein (1tul)	1 + 2	1 + 2	1 + 2	1 + 9
interleukin-1 β (1ilb)	21 + 27	21 + 26	21 + 26	21 + 31
α/β				
biotin carboxylase (1bnc)	2 + 0	2 + 4	2 + 0	2 + 7
heat-shock protein 70kDa (1atr)	17 + 2	17 + 2	17 + 2	17 + 2
thioredoxin (1ert)	24 + 0	24 + 0	24 + 0	24 + 69

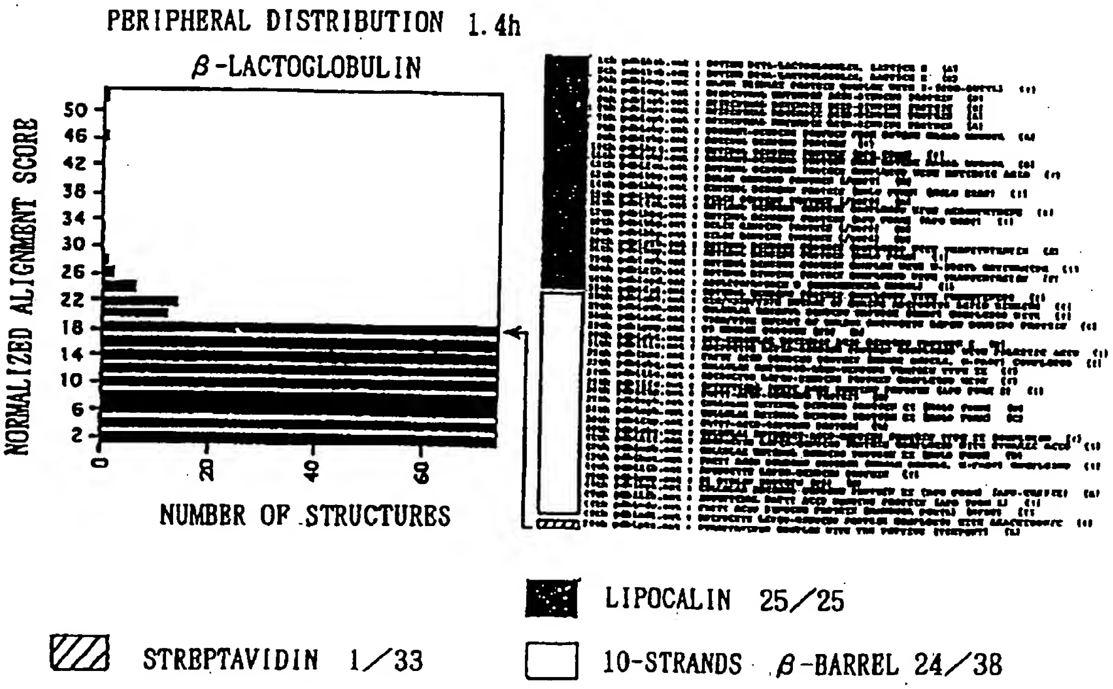
F I G. 12

PROTEIN	MARGINAL DISTRIBUTION	DDP	FAMILY+SIMILAR STRUCTURE
mainly α			
myoglobin	329 + 0	329 + 0	329 + 6
(1mbc)	2.1h	66.9h	–
cytochrome c	53 + 0	52 + 0	68 + 0
(1ccr)	1.4h	69.1h	–
cytochrome P450	31 + 0	–	31 + 0
(1oxa)	2.3h	–	–
mainly β			
β-lactoglobulin	25 + 25	25 + 0	25 + 71
(1beb)	1.4h	72.8h	–
telokin-like protein	1 + 2	–	1 + 9
(1tul)	1.2h	–	–
interleukin-1 β	21 + 26	–	21 + 31
(1ilb)	1.5h	–	–
α/β			
biotin carboxylase	2 + 4	–	2 + 7
(1bnc)	2.3h	–	–
heat-shock protein 70kDa	17 + 2	–	17 + 2
(1atr)	2.9h	–	–
thioredoxin	24 + 0	–	24 + 69
(1ert)	1.4h	–	–

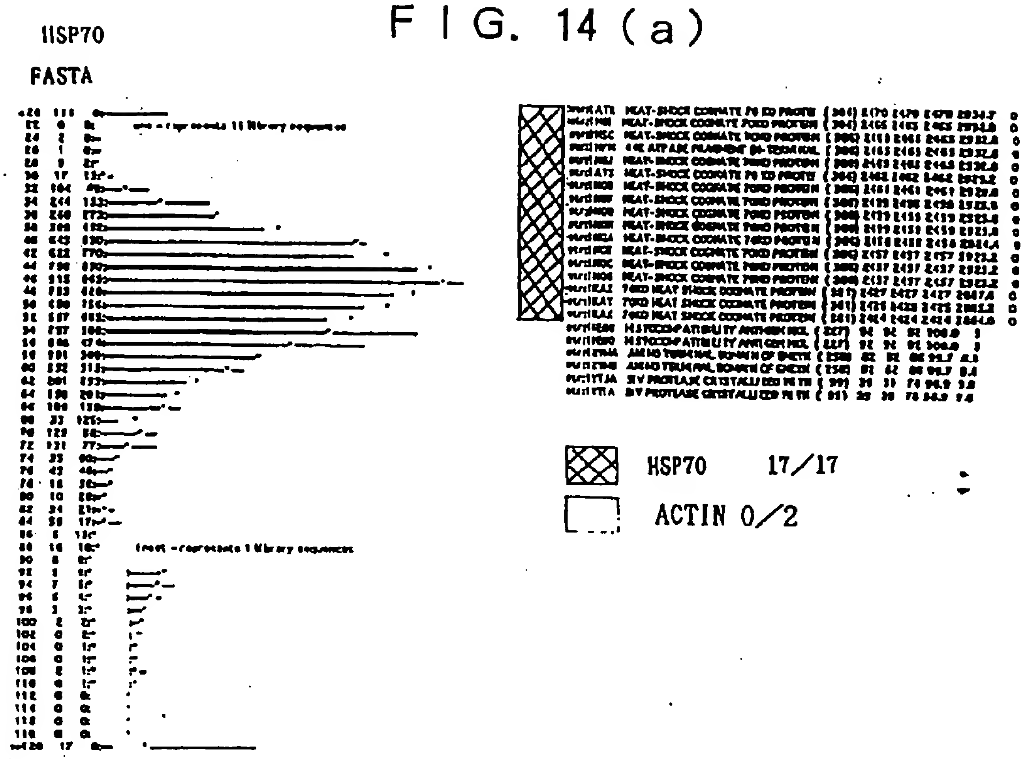
F I G. 13 (a)



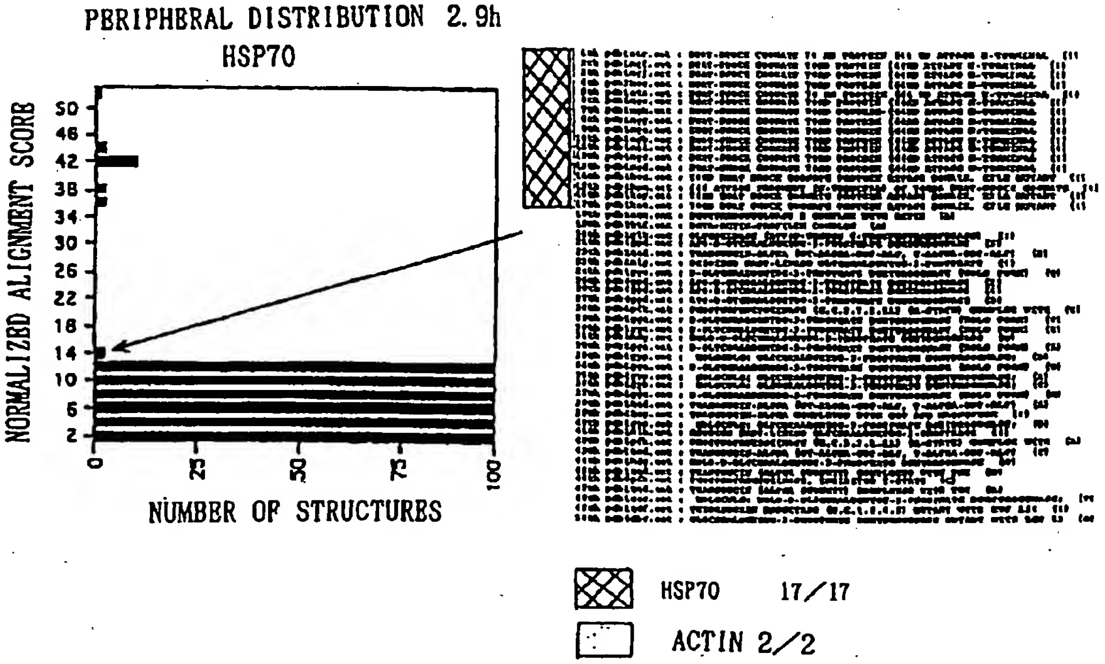
F I G. 13 (b)



F I G. 14 (a)

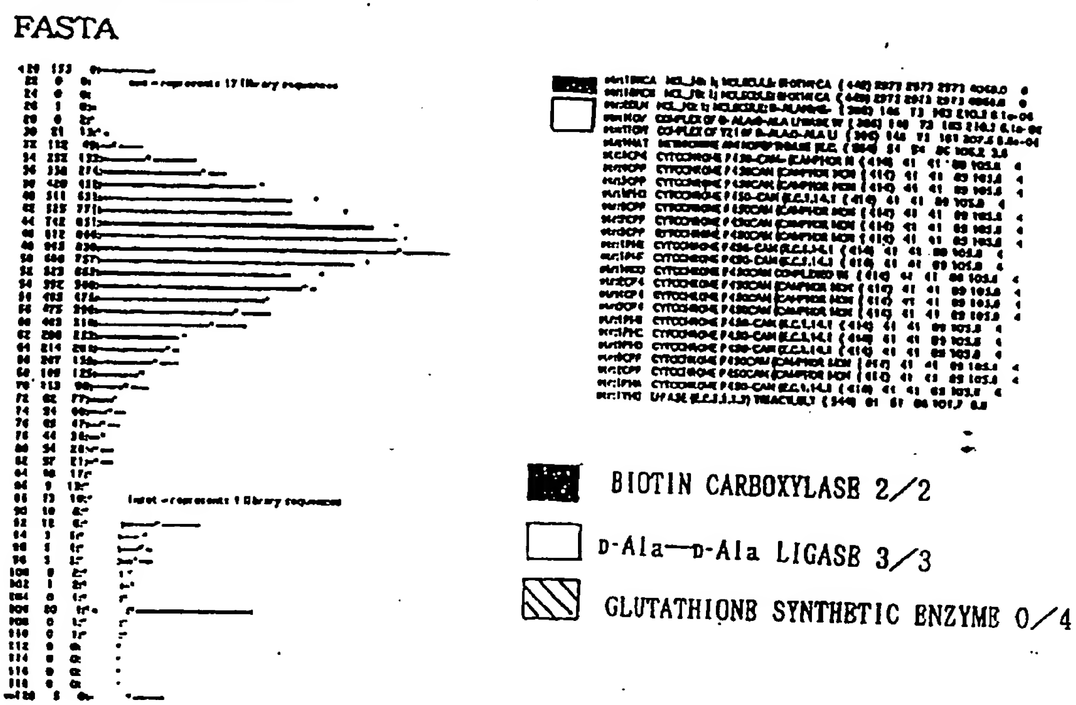


F I G. 14 (b)



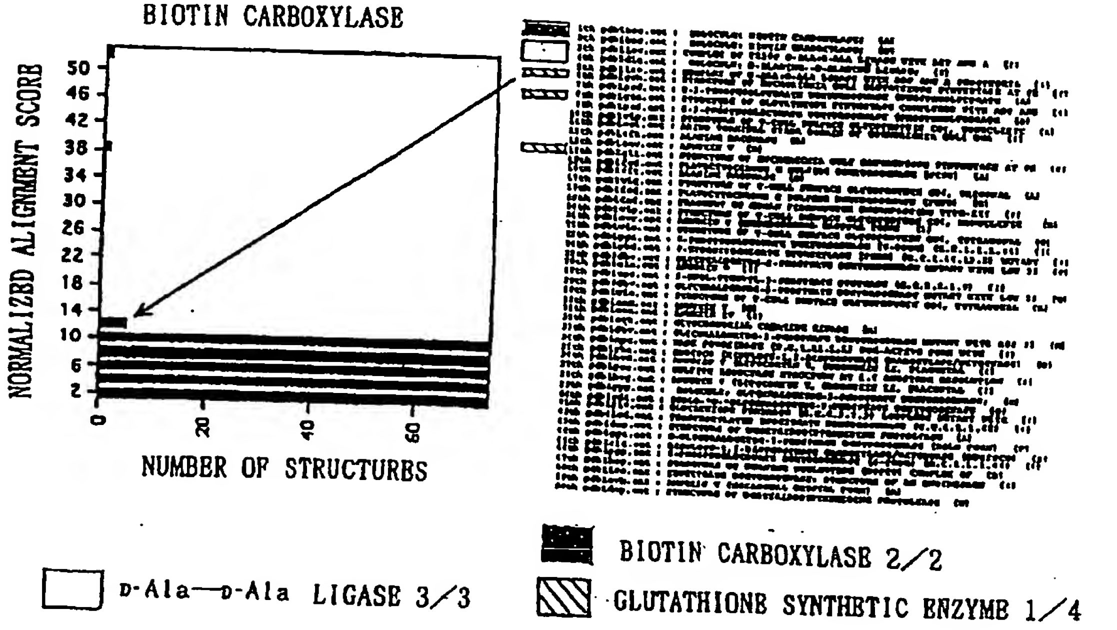
F I G. 15 (a)

BIOTIN CARBOXYLASE



F I G. 15 (b)

PERIPHERAL DISTRIBUTION 2.3h



European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 99 10 3018

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	PATENT ABSTRACTS OF JAPAN vol. 018, no. 322 (C-1214), 20 June 1994 & JP 06 073088 A (FUJITSU LTD), 15 March 1994 * abstract * ---	1	G06F17/30
A	PATENT ABSTRACTS OF JAPAN vol. 095, no. 006, 31 July 1995 & JP 07 056931 A (FUJITSU LTD), 3 March 1995 * abstract * ---	1	
A	PATENT ABSTRACTS OF JAPAN vol. 018, no. 522 (P-1808), 30 September 1994 & JP 06 180737 A (FUJITSU LTD), 28 June 1994 * abstract * ---	1	
A	PARISIEN M. ET AL.: "A Protein Conformational Search Space Defined by Secondary Structure Contacts" PROC. PACIFIC SYMPOSIUM ON BIOCOMPUTING, 4 - 9 January 1998, pages 425-436, XP002103078 Maui, Hawaii,USA * page 426, line 8 - page 431, line 9; figures 1-6 * -----	1	TECHNICAL FIELDS SEARCHED (Int.Cl.6) G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
BERLIN	19 May 1999	Deane, E

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document